JOURNAL OF LEARNING ANALYTICS

Volume 12(1), 12–31. https://doi.org/10.18608/jla.2025.8591

Mapping the Landscape of Generative Artificial Intelligence in Learning Analytics: A Systematic Literature Review

Kamila Misiejuk[1], Sonsoles López-Pernas[2], Rogers Kaliisa[3] and Mohammed Saqr[4]

Abstract

Generative artificial intelligence (GenAI) has opened new possibilities for designing learning analytics (LA) tools, gaining new insights about student learning processes and their environment, and supporting teachers in assessing and monitoring students. This systematic literature review maps the empirical research of 41 papers utilizing GenAI and LA and interprets the results through the lens of the LA/EDM process cycle. Currently, GenAI is mostly implemented to automate discourse coding, scoring, or classification tasks. Few papers used GenAI to generate data or to summarize text. Classroom integrations of GenAI and LA mostly explore facilitating human–GenAI collaboration, rather than implementing automated feedback generation or GenAI-powered learning analytics dashboards. Most papers use Generative Adversarial Network models to generate synthetic data, BERT models for classification or prediction tasks, BERT or GPT models for discourse coding, and GPT models for tool integration. Although most studies evaluate the GenAI output, we found examples of using GenAI without the output validation, especially when its output feeds into an LA pipeline aiming to, for example, develop a dashboard. This review offers a comprehensive overview of the field to aid LA researchers in the design of research studies and a contribution to establishing best practices to integrate GenAI and LA.

Notes for Practice

- While GenAI models like BERT and GPT can significantly reduce the time and effort required by LA researchers for analyzing textual data, continuous improvement in models is needed.
- GAN algorithms do not always perform as well as other approaches, but they create more realistic data and dominate the field of data generation.
- LA researchers should explore the use of GenAI in different educational contexts, develop standardized evaluation metrics for GenAI applications, and establish best practices for reliable integration of these technologies into the LA pipeline.
- Evaluation studies showed positive student perceptions of GenAI, with improvements in task performance and participation. More research is needed on its impact on actual learning outcomes.

Keywords: Generative artificial intelligence, GenAI, learning analytics, educational data mining, systematic review

Submitted: 04/08/2024 — **Accepted:** 20/02/2025 — **Published:** 20/03/2025

Corresponding author [1]Email: kamila.misiejuk@fernuni-hagen.de Address: Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA), FernUniversität in Hagen, Universitätsstrasse 27, 58097 Hagen, Germany. ORCID iD: https://orcid.org/0000-0003-0761-8703

[2]Email: sonsoles.lopez@uef.fi Address: School of Computing, University of Eastern Finland, Joensuu Campus, Yliopistokatu 2, FI-80100 Joensuu, Finland. ORCID iD: https://orcid.org/0000-0002-9621-1392

[3]Email: rogers.kaliisa@iped.uio.no Address: Department of Education, the University of Oslo, Post Box 1092, Blindern 0317, Oslo, Norway. ORCID iD: https://orcid.org/0000-0001-6528-8517

[4]Email: mohammed.saqr@uef.fi Address: School of Computing, University of Eastern Finland, Joensuu Campus Yliopistokatu 2, FI-80100 Joensuu, Finland. ORCID iD: https://orcid.org/0000-0001-5881-3109

1. Introduction

For over a century, humans have aspired to harness machine intelligence to improve teaching and learning (Pressey, 1926). The history includes machines that would automate the generation of questions to students (Pressey, 1926); "teaching machines" that present and automatically grade questions (Skinner, 1958), and computer-assisted instruction where the

computers would deliver learning materials adapted to student needs (Suppes, 1966). In tandem with the evolution — and recurring surges in interest — in artificial intelligence (AI), several tutoring technologies have been created and advanced to incorporate AI in education (AIED; McCalla, 2023). Nonetheless, the road to today's AI was paved with spells of hype, setbacks, and disillusionment, often referred to as AI winter (periods of skepticism and decline of research; Perrotta & Selwyn, 2020). While it is unclear if AI winter has ever curtailed research on AI in education, it is rather evident that the latest surge in AI was coupled with the emergence of data-intensive educational fields, i.e., learning analytics (LA) and educational data mining (EDM; McCalla, 2023; Romero & Ventura, 2020). Driven by digitization, an abundance of data, and burgeoning computing power, the three fields (AIED, EDM, and LA) have grown exponentially across several domains of research and practice. The fact that the three fields share interdisciplinary roots across computing, cognitive, and social sciences has given rise to common — and often overlapping — perspectives and applications. Recently, the striking development of generative AI (GenAI) capabilities has sent ripples of enthusiasm — and fear — across the world. For the first time, AI can perform high cognitive functions that were a unique preserve of well-educated humans (Noy & Zhang, 2023). A new AI reality has ensued, along with another surge of interest, enthusiasm, and hype. Harnessing the capabilities of GenAI has been a top priority and a thematic objective of researchers, institutions, and governments alike. In this paper, we examine how GenAI has been used, integrated, or used to augment LA or EDM to improve teaching and learning.

2. Background and Related Work

The integration of GenAI into educational technologies has resulted in several possibilities for LA. Capitalizing on the capabilities of GenAI, researchers and practitioners could potentially design sophisticated LA tools that offer personalized learning experiences in real time. For instance, augmenting educational decisions through GenAI-assisted feedback and advisory systems supported by advanced algorithms and LA techniques could offer timely, personalized, data-driven feedback to teachers and students (Lodge et al., 2023). This real-time responsiveness could enhance the learning experience by providing teachers and students with actionable insights that could be immediately incorporated into their learning design and study strategies (González-Calatayud et al., 2021). The goal of many recently developed AI-powered systems was to support teachers with data-driven decisions to improve their practice, decrease their workload, and organize their classrooms more effectively. If implemented effectively, GenAI technologies have the potential to streamline the grading process and thus reduce educator workloads and provide more accurate and consistent formative assessment and feedback provision (Hopfenbeck et al., 2023) by providing instructional suggestions to teachers for them to adopt or ignore (Luckin et al., 2022). At the same time, proper integration of GenAI tools requires students and instructors to develop new skills, such as prompting (Misiejuk, López-Pernas, et al., 2024) or GenAI literacy (Bozkurt, 2024).

2.1. Related Work

Recent reviews have highlighted the breadth of GenAI's impact while also identifying gaps and future research directions. Kumar et al. (2023) offer a comprehensive overview of large GenAI models like ChatGPT, DALL-E 2, and Stable Diffusion. They categorize these models based on their input and output formats and discuss their potential to impact industries by automating and enhancing various tasks. However, their focus is primarily on the capabilities and applications of these models, with limited discussion on the technical mechanisms or ethical considerations. Schöbel et al. (2024) map the evolution and future of conversational agents (CAs) through a bibliometric analysis of over 5,000 research articles. They identify the "five waves" of CA research, detailing the technological advancements and theoretical paradigms from early chatbots to modern generative models like GPT-3 and BLOOM NLU. This study emphasizes the continuous technological advancements in CAs and their impact on various domains but does not extensively cover other applications of GenAI, such as LA. Sengar et al. (2024) document the systematic review and analysis of recent advancements and techniques in GenAI with a detailed discussion of their applications, including application-specific models. García-Peñalvo and Vázquez-Ingelmo (2023) reviewed 631 GenA1 studies to characterize the GenAI landscape. The study identifies a dichotomy in the understanding and application of GenAI, highlighting the difference between public perception and the AI research community's focus. Gupta et al. (2024) used topic modelling to analyze 1,319 studies on GenAI between 1985 and 2023. The study identified seven clusters of topics, including image processing, content generation, emerging use cases, engineering, cognitive inference and planning, data privacy and security, and GPT applications.

In the educational context, Guo et al. (2024) provided a bibliometric analysis of AIED research, examining 6,843 publications from 2013 to 2023. Their study highlights key research trends, showing that AI applications in education predominantly focus on higher education and STEM fields, with growing interest in intelligent tutoring systems, automated grading, and recommender systems. However, their review notes a lack of research on AI integration in K-12 and preschool education, as well as limited discussion on the ethical and pedagogical implications of AI technologies in teaching and learning. More recently, Yan et al. (2024) have delved into the integration of GenAI within the LA/EDM cycle. They contextualize the opportunities and challenges of GenAI within Clow's LA cycle framework, highlighting its potential in analyzing unstructured data, generating synthetic learner data, enriching multimodal interactions, advancing interactive analytics, and facilitating

personalized interventions. The review also emphasizes the need for a renewed understanding of learners in the age of GenAI and advocates for frameworks that support human–AI collaboration in learning contexts. As well, Ouyang and Zhang (2024) examined AI-driven learning analytics applications and tools within the context of computer-supported collaborative learning (CSCL) and analyzed 26 studies from a pool of 2,607 articles. The findings reveal a primary focus on cognitive engagement and the use of communicative discourse, behavioural, and evaluation data for feedback and visualization, with a notable lack of design principles guiding tool development and insufficient use of AI for instructional interventions.

2.2. Motivation, Aims, and Research Questions

While there is a substantial body of research on the application and potential of GenAI, to the best of our knowledge, to date, there is no systematic review mapping the existing integrations of GenAI in the LA pipeline, including the distinct ways AI and LA have been intertwined. The closest study to ours is Yan et al. (2024), which mapped the opportunities and challenges of GenAI in the LA/EDM cycle. While their study offers valuable insights into the intersection of GenAI and LA, our research diverges by focusing specifically on the empirical mapping of how GenAI is currently employed across various stages of the LA/EDM cycle. Our systematic review of empirical studies aims to provide a more granular analysis of the practical implementations, methodological approaches, and outcomes associated with GenAI in LA. This study not only complements the theoretical foundation laid by Yan et al. (2024) but also further contributes to the literature by offering concrete examples from each LA/EDM cycle and developing a new framework to map the integration of LA and GenAI. We interpret the findings through the lens of the LA/EDM process cycle characterized by the dimensions of *data capture, data pre-processing and preparation, analysis and integration, insightful action, and feedback* (Saqr, 2018). Thus, we provide a structured overview of how GenAI is transforming each step of the LA pipeline. The insights and implications drawn offer a roadmap for future research, pointing to areas where GenAI can further contribute to LA research and vice versa. The study will address the following research questions:

RQ1: How does GenAI impact/contribute to the LA process cycle, from data to evaluation?

The LA cycle conceptualizes successful LA work as four linked steps: 1) learners 2) generating data 3) used to produce metrics, analytics, or visualizations, and 4) "closing the loop" by feeding back this product to learners and teachers through one or more interventions (Clow, 2012). The first research question explores how GenAI contributes to the LA/EDM process cycle through the automation of complex tasks and the provision of advanced analytics capabilities (Clow, 2012). For example, during data pre-processing and preparation, GenAI could be used to automate data cleaning and coding to ease this task for educators and researchers who may lack the technical expertise to handle challenging data management tasks. GenAI models have the potential to perform advanced analyses, discovering patterns and correlations that traditional methods might miss (Yan et al., 2024).

RQ2: What specific problems in LA research are addressed by GenAI? What solutions have been proposed or implemented, and how?

Some of the critical challenges highlighted across the LA literature include the difficulty in managing unstructured data (Zhan et al., 2023), data privacy, and teachers' limited technological expertise in making sense of the variety of actions in terms of visualizations and textual feedback provided by LA tools (Kaliisa et al., 2022; Ferguson, 2019). This research question explores whether the integration of GenAI can address some of the challenges faced by LA researchers and practitioners and identify the specific ways this is achieved (Ferguson, 2019). For example, GenAI could be used to generate synthetic data, mitigating the risks associated with using real student data (Yan et al., 2024) while maintaining utility for LA models (Dorodchi et al., 2019). Additionally, GenAI might enhance personalization in learning by creating adaptive learning paths tailored to individual student needs and specific pedagogical frameworks (Kaliisa et al., 2022). Thus, GenAI has the potential to address key challenges in LA through the improvement of LA tools while also ensuring ethical compliance and data privacy, making it a valuable addition to the field of LA.

Furthermore, understanding the actions or interventions that follow LA analysis integrating GenAI is essential because it highlights the practical applications and impacts of these technologies in educational settings. One major challenge in LA is the alignment of LA outputs with teaching practices (Kaliisa et al., 2022). In this study, we investigate whether GenAI could bridge this gap by providing actionable insights directly linked to specific pedagogical strategies (Clow, 2012). For instance, GenAI could support adaptive learning by dynamically adjusting the learning content and activities based on individual student needs, ensuring timely and tailored interventions. Additionally, GenAI-driven dashboards could offer real-time recommendations for instructional adjustments, allowing educators to respond promptly to student needs. However, there is a lack of evidence to prove these claims about GenAI's impact on LA interventions, thus being a focus for the current study.

3. Methods

This systematic literature review follows the Preferred Reporting Items of Systematic Review and Meta-Analysis (PRISMA) framework (Page et al., 2021). On the 5th of July 2024, we searched three scientific databases, Scopus, Web of Science, and ERIC, using the following search string: *("learning analytics" OR "educational data mining") AND (ai OR llm OR generative*

OR prompt OR "artificial intelligence"). The search resulted in a total of 1,786 papers, including 995 papers from Scopus, 394 papers from Web of Science, and 397 papers from ERIC (see Figure 1). The papers were exported to a web-based review software, Rayyan-ai (Johnson & Phillips, 2018). After removing 392 duplicates, the remaining 1,394 papers were collaboratively sorted by three researchers based on their titles and abstracts according to their relevance to the research questions and quality. The included studies had to fulfill the following criteria: 1) be peer-reviewed empirical studies, 2) be situated in an educational setting, 3) be written in English, and 4) use GenAI methods, e.g., large language models (LLMs) or Generative Adversarial Networks (GANs). The first round of screening resulted in the exclusion of 1,319 papers and the inclusion of 75 papers for full-text reading.

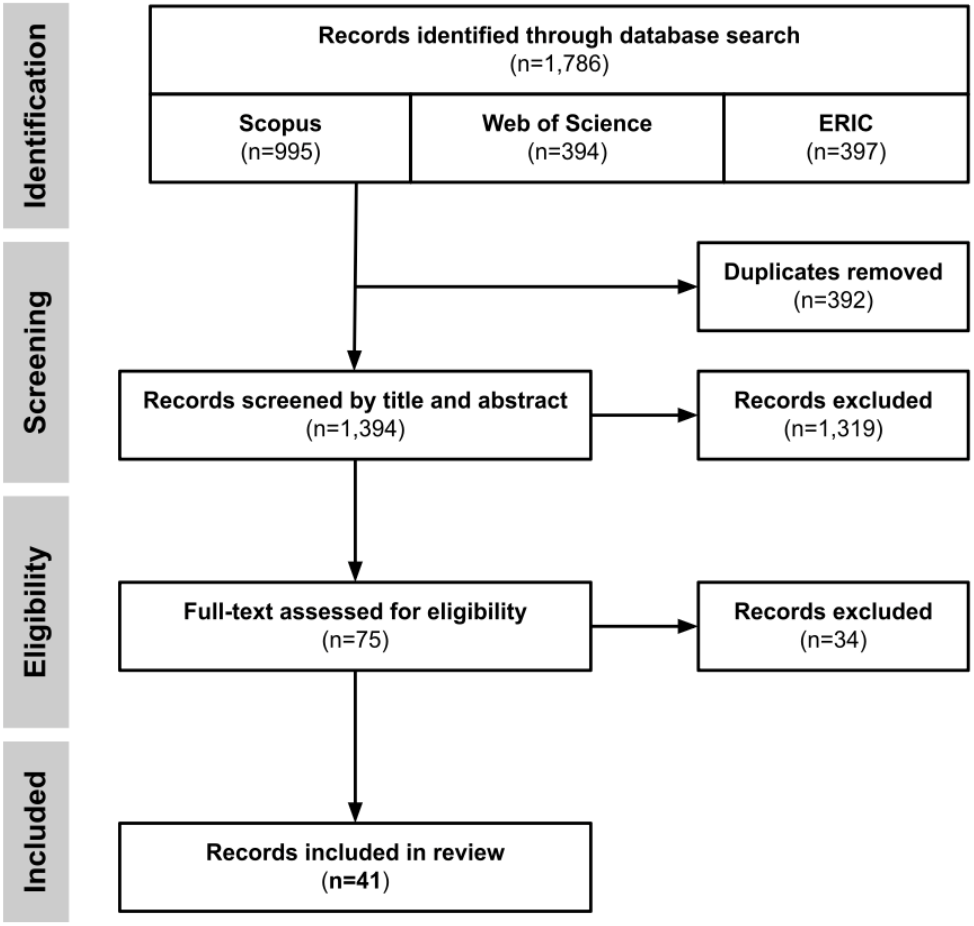

Figure 1. PRISMA flow diagram.

The second round of the review included full-text reading of papers, quality assessment, and mapping the use of GenAI in each paper on the LA/EDM process cycle. As such, we excluded all papers that did not use LA or EDM approaches and verified if all papers did use GenAI. This led to the exclusion of 34 papers and the total final dataset of 41 papers.

We used the framework contextualizing the implications of GenAI within each step of the LA cycle by Yan et al. (2024) and the LA/EDM cycle framework by Saqr (2018) for the initial mapping of the papers. This mapping was expanded by the insights from the full-text reading of the papers and led to the development of the use of GenAI in the LA cycle framework (see Figure 2). The goal of this framework is to examine the purpose of each GenAI implementation at a specific stage in the LA/EDM cycle. In addition, it maps the intertwined nature of LA-GenAI research with different degrees of integration, where GenAI can be used to analyze LA data, as well as LA can be used to analyze GenAI. The framework includes five stages:

1. ***Data*** refers to the types of data generated by GenAI or in collaboration with GenAI for LA analysis.
2. ***Pre-processing and preparation*** is defined as a stage in which GenAI is used to pre-process or prepare data for LA analysis.
3. ***Methods, analysis, and tools*** includes using GenAI as a method to analyze LA data or development of GenAI-powered tools.
4. ***Implementation and action*** refers to implementations of GenAI and LA in a classroom.
5. ***Evaluation*** indicates the use of LA to examine the integration of GenAI to provide insights into future implementations.

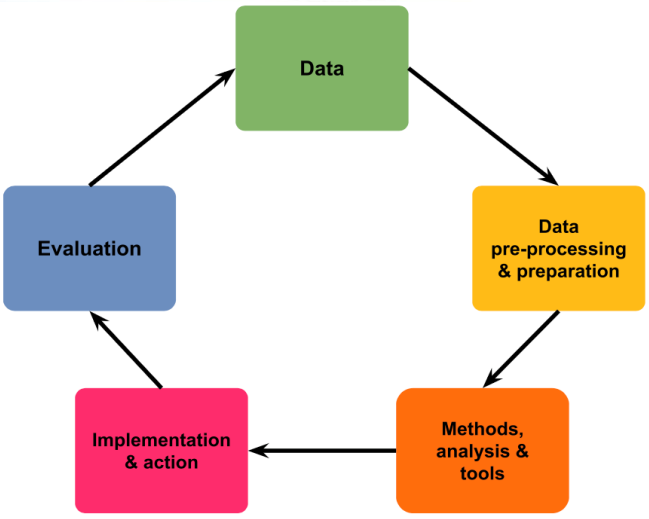

Figure 2. The use of GenAI in the LA/EDM cycle framework.

After classifying all implementations of GenAI in each stage, we extracted the following details from the research papers to answer our research questions: 1) *The role of GenAI*: the issue that the use of GenAI is intended to address; 2) *Overall LA task*: the overall research aim of the research; 3) *GenAI technology*: the technology implemented to solve a specific challenge; 4) *Data type/sources*: data used by the GenAI implementation; 5) *Evaluation method*: the metrics and validation method(s) that evaluate the GenAI performance; 6) *Results:* results of the GenAI evaluation for a specific task.

4. Results

The results of this systematic review mapped the use of GenAI in the LA/EDM cycle framework (see Figure 3). The following sections describe the GenAI implementation at each stage of the LA/EDM cycle in detail. Our dataset includes one paper published in 2020, five papers published in 2021, six papers published in 2022, 11 papers published in 2023, and 18 papers from 2024. The list of all papers included in this literature review can be found in Appendix A.

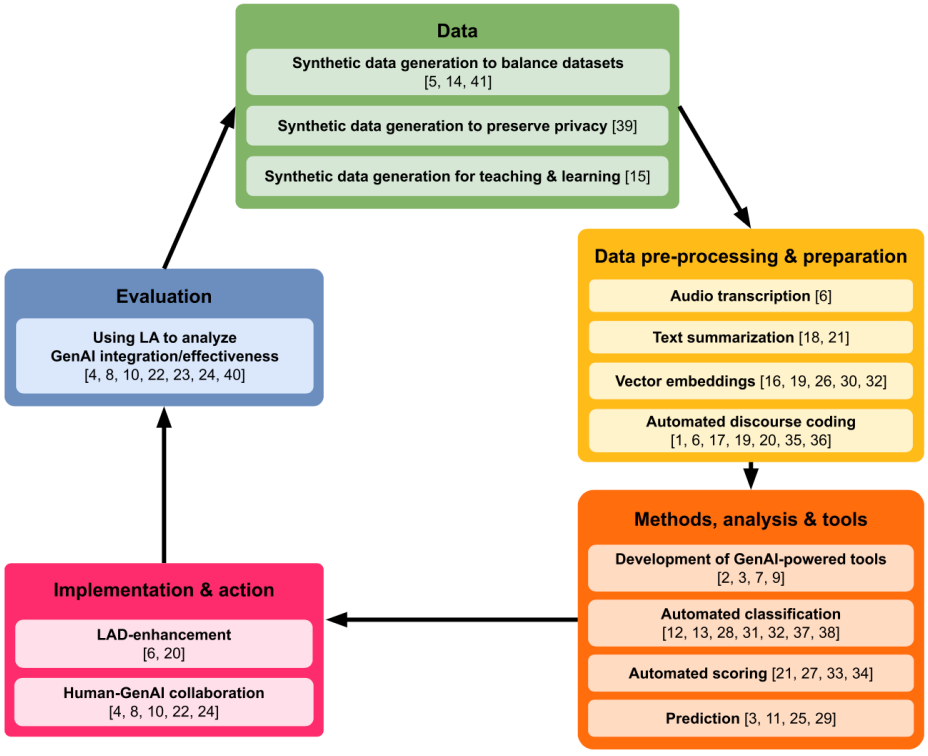

Figure 3. Mapping of the empirical papers on the use of GenAI in the LA/EDM cycle framework.

Note: *Here we use the number assigned to each paper, which can be found in the full list of 41 papers we analyzed; however, the list is in alphabetical rather than numerical order.*

4.1. Data

The *data* category encompasses articles that made use of GenAI for its data generation capabilities. A total of five papers were included in this category (see Table 1). All of which proposed the use of some variation of GANs. Four studies (Vives et al., 2024; Wu et al., 2022; Q. Liu et al., 2024; Ying & Ma, 2024) used GenAI for the generation of synthetic data to carry out one of the most common LA tasks: performance prediction. The datasets included data on student demographic information, behavioural data, and performance indicators. Three studies (Vives et al., 2024; Volarić, et al., 2023; Ying & Ma, 2024) operationalized synthetic data to **augment an existing dataset**, with the goal of overcoming the problems of unbalanced data that often arise when predicting unlikely phenomena such as dropouts. In turn, Q. Liu et al. (2024) used synthetic data generation to create completely new datasets to perform prediction tasks while **preserving student privacy**. Researchers compared the performance of GenAI versus other synthetic data generation algorithms to the original dataset across various prediction models. In general, the results show that GANs can generate realistic data (Vives et al., 2024) or preserving privacy (Q. Liu et al., 2024) while having comparable performance to the original data (Volarić et al., 2023; Ying & Ma, 2024). Other synthetic data generation algorithms outperform GANs in some tasks, such as SMOTE (Vives et al., 2024) or Gaussian Multivariate (Q. Liu et al., 2024). The remaining article in this category addresses a completely different topic, which is the **generation of exam questions** from a pool of existing questions. The GAN-generated questions were evaluated by teachers in terms of difficulty, distinguishability, rationality, and validity, and outperformed those generated by other algorithms.

Table 1. Reviewed Articles in the *Data* Category

Ref.	Role of GenAI	Overall LA task	GenAI technology	Data type/sources	Evaluation method	Results
Vives et al., 2024	Synthetic data generation to balance datasets	Performance prediction	GAN	Programming grades, linguistic comprehension, and mathematics scores	Performance comparison to baseline and SMOTE using 6 algorithms and performance metrics	SMOTE more balanced; GAN more realistic
Volarić et al., 2023	Synthetic data generation to balance datasets	Performance prediction	GAN	OULAD: VLE data (assessment, clicks...)	Performance comparison to baseline using 5 algorithms and performance metrics	Comparable; GAN balanced the dataset
Q. Liu et al., 2024	Synthetic data generation to preserve privacy	Performance prediction	CTGAN	3 datasets with demographic, behaviour, and performance data	Performance: Comparison to baseline, GM and GC using performance metrics Resemblance evaluation: Difference in pairwise correlation, JS, WD Privacy evaluation: DCR, NNDR, MIA	GM best performing; CTGAN best resemblance and privacy evaluation
Ying & Ma, 2024	Synthetic data generation to balance datasets	Performance prediction	GAN	Performance in Mathematics and Portuguese	Performance comparison to baseline using 3 algorithms and performance metrics	Comparable for RF and SVR; increased performance for NN
Wu et al., 2022	Synthetic data generation for teaching and learning	Exam generation	Conditional GAN	3 datasets: 2 datasets are from ASSISTments, a math electronic tutor; 1 is from an engineering course	Performance comparison to RSF, GA, Conditional VAEs (ExamVAE) in terms of difficulty, distinguishability, rationality, validity	GAN best performing

VLE: Virtual Learning Environment; SMOTE: Synthetic Minority Over-Sampling Technique; CTGAN: Conditional Tabular GAN; RF: Random Forest; SVR: Support Vector Regression; NN: Neural Network; RSF: Random Sampling and Filtering; GA: Genetic Algorithm; VAE: Variational Autoencoder; JSD: Jensen-Shannon Divergence; WD: Wasserstein Distance; DCR: Distance to Closest Record; NNDR: Nearest Neighbour Distance Ratio; MIA: Membership Interface Attack; GM: Gaussian Multivariate; GC: Gaussian Copula.

4.2. Pre-processing and Preparation

Data pre-processing and preparation papers use GenAI to pre-process or prepare data for LA analysis. This category includes ten papers with GenAI implementations mapped into four categories: audio transcription, vector embeddings, automated data coding, and text summarization (see Table 2).

Only one paper used GenAI for **audio transcription**. Milesi et al. (2024) used GenAI to transcribe audio recordings containing nurses' dialogues during teamwork simulations using OpenAI Whisper. These recordings were later automatically coded to identify critical communication using GenAI and the coded data was used to develop a learning analytics dashboard (LAD).

Overall, seven papers (Garg et al., 2024; Milesi et al., 2024; Pan et al., 2020; Menzel et al., 2023; Pinargote et al., 2024; Pugh et al., 2021; Misiejuk, Kaliisa, & Scianna, 2024) used a variety of BERT and GPT models to **automatically code discourse data**, such as transcripts of collaborative sessions or discussion forum posts. Three papers (Garg et al., 2024; Pugh

et al., 2021; Misiejuk, Kaliisa, & Scianna, 2024) were focused on evaluating the accuracy of coding and evaluating effective prompting strategies with different GenAI models. Four papers integrated GenAI into their LA pipeline and adopted the GenAI output into their LAD development (Pan et al., 2020; Milesi et al., 2024; Pinargote et al., 2024). For example, Pan et al. (2020) analyzed students' scientific argumentation by GenAI and time-series data using Hidden Markov Models (HMMs) to identify their cognitive state transition. The result visualizations are currently being integrated into a LAD. In addition, the correlation analysis using the GenAI-coded data showed a positive relationship between student learning performance and high-level scientific argumentation skills, as well as a positive relationship between correct solutions and flow state and that correct solutions are negatively correlated with anxiety and boredom states. In contrast, Menzel et al. (2023) used the GenAI-generated text data coding to cluster students based on their learning patterns and provide them with feedback specific to their cluster group. Only two studies (Milesi et al., 2024; Menzel et al., 2023) did not evaluate the coding output of GenAI algorithms. One study (Pugh et al., 2021) compared GenAI coding with another automated method, RF, while other studies (Garg et al., 2024; Pinargote et al., 2024; Misiejuk, Kaliisa, & Scianna, 2024) focused on assessing human–GenAI reliability using several metrics, such as Cohen's κ or F1-score. The evaluation results showed that GenAI algorithms perform better than Random Forest; the human–GenAI agreement is still substantial at best.

Although **vectorization** is typically part of BERT implementation, one study (Menzel et al., 2023) specifically decided to use BERT over a commonly used Latent Semantic Analysis to create vectors from discourse forum data as part of the Group Communication Analysis method. Another study (C.-C. Lin et al., 2024) converted log data into natural language sentences that were later transformed into semester-based or daily embeddings using three different approaches: Gemini, BERT, and OpenAI embeddings methodology. The embeddings were used to predict student performance. The study reported that BERT was best suited to model daily learning activities, while OpenAI embeddings performed best for semester-based predictions. Sha et al. (2021) used BERT for word embedding in two out of five models that they used for an automated classification task. In three other models, BERT was coupled only with a classification layer and fine-tuned. The models using BERT for fine-tuning showed better performance than models using BERT for word embeddings; however, the performance improvement was limited. Two papers in this group examined the input effectiveness of BERT models. Wu et al. (2023) used factor analysis to examine which input should be part of a prompt, including a combination of student answers, correct definitions, or the word being asked. The results showed that including the student's answer and the correct definition/sentence had the best separability; thus, it was used for the hyperparameter tuning of the model. At the same time, Wu et al. (2023) reported that human–AI agreement was lower than human–human agreement. In a similar study, Condor et al. (2021) examined the following potential input content: question text, question context, rubric text, or bundle identifier. Different combinations of prompt input were evaluated for SBERT, Word2Vec, and Bag-of-words methods using multinomial logistic regression and a neural network classification model. The human rating was chosen as the ground truth. It was found that SBERT representations performed best, while the inclusion of question text or question context improved model performance.

Finally, two studies (Snyder et al., 2024; Yang et al., 2021) used GenAI to **summarize text**. The BERT approach outperformed two traditional NLP methods, TextRank and RAKE in Yang et al. (2021) in summarizing e-book text markings. BLEU score — which measures the similarity of the machine-translated text to a set of reference translations — and the METEOR score — which assesses the test using stemming, synonymy matching, and standard exact word matching — were used to evaluate the summarization output. The GenAI-generated summaries were used to calculate the BLEU score of student marking every week during a semester, and the sum of scores constituted students' final grades. This study also collected other log data and conducted a survey on self-regulated learning skills. Students were clustered into groups based on their grades and marking frequencies. The results showed, among others, that high-skill readers who prefer marking have better learning performance, task strategies, and time management than high-skill readers who do not like marking. The summarizations of collaborative problem-solving sessions by GPT-3.5 Turbo in Snyder et al. (2024) were evaluated by humans hand-coding the GenAI output. However, the authors report encountering hallucination issues or that the LLM would refuse to provide the summary due to a lack of confidence when developing the summarization prompt. The results of the human coding of GenAI summarizations were analyzed using Markov Chain analysis to identify sequences of collaborative problem-solving behaviours. Distinct behavioural patterns were identified. High-performing students utilized more transitions from planning to enacting states, as well as more effective collaboration strategies, such as sharing plans within the group before enacting. In comparison, the transitions of low-performing students were more disjointed and included fewer discussions among the collaborating students.

Table 2. Reviewed Articles in the *Pre-processing and Preparation* Category

Ref.	Role of GenAI	Overall LA task	GenAI technology	Data type/sources	Evaluation method	Results
Milesi et al., 2024	**Audio transcription**	LAD development	OpenAI Whisper	Audio recordings of teamwork simulations	-	-

Pan et al., 2020	**Automated discourse coding**	LAD development	BERT	Students' scientific argumentation	Comparison with RF; Baseline: Human coding; Evaluation metrics: A, F1, MCC	BERT performed best
Pugh et al., 2021	**Automated discourse coding**	Automated transcription accuracy	BERT, BERT-seq	Transcripts of collaborative problem-solving sessions	Comparison with RF; Evaluation metrics: AUROC	BERT-seq performed best, but results not statistically significant for all codes
Pinargote et al., 2024	**Automated discourse coding**	LAD development	ChatGPT-3.5	Transcripts of collaborative sessions	Comparison with human coding; Evaluation metrics: Krippendorff's Alpha (human–human), Cohen's κ (human–AI)	Highest human–AI agreement level: *substantial*
Milesi et al., 2024	**Automated discourse coding**	LAD development	BERT	Transcripts of teamwork simulations	-	-
Garg et al., 2024	**Automated discourse coding**	LA method development	GPT-3.5–Turbo, GPT-4	Teacher interviews	Comparison of 4 prompting methods and fine-tuning; Baseline: Human coding; Evaluation metrics: Shaffer's Rho, F1, Cohen's κ	GPT-3.5–Turbo fine-tuning performed best, but no approach achieved sufficient reliability
Menzel et al., 2023	**Automated discourse coding**	Automated feedback generation	RoBERTa	Discussion forum posts	-	-
Misiejul, Kaliisa, et al., 2024	**Automated discourse coding**	LA method development	GPT-4 and text-davinci-003 API models, ChatGPT-4 Code Interpreter	Discussion forum posts	Comparison with human coding; Evaluation metrics: A, P, R, F1, Cohen's κ, Gwet AC1, MCC	GPT-4 API model performed best; Highest human–AI agreement achieved: *substantial*
Menzel et al., 2023	**Vector embeddings**	Automated feedback generation	SBERT	Discussion forum posts	-	-
C.-C. Lin et al., 2024	**Vector embeddings**	Performance prediction	BERT, Gemini: models/embedding-001, OpenAI: text-embedding–3-large	BookRoll and Viscode system logs converted into natural language sentences	Developed two predictive models with datasets built using different embedding methods; Evaluation metrics: F1, R, P, A; T-SNE clusters and visualization scores	OpenAI best for semester embeddings; BERT best for daily embeddings
Srivastav et al., 2024	**Vector embeddings**	Forum post classification	BERT	Forum posts	Comparison with NB, SVM, RF, LR and models using BERT for fine-tuning; Evaluation metrics: A, Cohen's κ, AUC, F1	BERT used at the classification layer performed better than used at the embedding layer
Condor et al., 2021	**Vector embeddings**	Automatic short answer grading	SBERT	Student responses to Critical Reasoning for College Readiness test	Comparison with Word2Vec and Bag-of-words (different combinations of content to vectorize and concatenate to the response vectors); Baseline: Human coding; Evaluation metric: Generalizability, Leave-one-bundle-out metric, A, F1	SBERT performed best; the addition of the question text or question context improved the generalizability of the model
Wu et al., 2023	**Vector embeddings**	Vocabulary acquisition	BERT	Transcription of student free-form spoken responses	Comparison of different prompt variations; Baseline: Human coding; Evaluation metric: Quadratic weighted kappa	Student's answer and the correct definition/sentence should be provided in a prompt for classifying definitions and sentence usage; human–AI agreement lower than human–human

Yang et al., 2021	Text summarization	Modelling e-book text marking skills	BERT	Learning materials	Comparison with TextRank and RAKE; Baseline: Instructor marking; Evaluation metrics: BLEU and METEOR scores	BERT performed best except one metric (BLEU-4)
Snyder et al., 2024	Text summarization	Modelling collaborative problem solving	GPT-3.5–Turbo	Transcripts of collaborative problem-solving sessions	Summarizations hand-coded by humans	-

A: Accuracy; AUROC: Area under the ROC curve; MCC: Matthew's Correlation Coefficient; P: Precision; R: Recall; RF: Random Forest.

4.3. Methods, Analysis, and Tools

Twenty papers use ***methods, analysis, or tools***, in which GenAI is used to analyze LA data or for the development of GenAI-powered tools. The applications of GenAI in this category are divided into four groups: automated classification, automated scoring, prediction, and tool development (see Table 3).

Seven papers (Z. Liu et al., 2023; Hu et al., 2024; Samadi et al., 2024; Jia et al., 2021; Sha et al., 2021; J. Lee & Koh, 2023; Y. Li et al., 2022) used GenAI for **automated classification**. Five papers (Z. Liu et al., 2023; Hu et al., 2024; Samadi et al., 2024; Sha et al., 2021; J. Lee & Koh, 2023) built classifiers for student interaction in discussion forums or chat, one paper (Jia et al., 2021) developed a classifier for peer review comments, while another paper (Y. Li et al., 2022) used learning objectives. All papers implemented a variation of a BERT model for classification. Only one paper (Samadi et al., 2024) examined GPT-2 models for classification purposes. All studies evaluated the GenAI classifier's performance by comparing it with the performance of other automated classifiers. Typically, human coding was established as a ground truth of the evaluation. F1-score was the most popular evaluation metric, found in all papers, followed by accuracy used in four papers (Sha et al., 2021; Samadi et al., 2024; Z. Liu et al., 2023; Y. Li et al., 2022). BERT models were found to outperform other machine learning models, such as Random Forest, Naïve Bayes, or Recurrent Neural Network in all studies. In addition, Jia et al. (2021) demonstrated that using a multi-task learning (MTL) BERT model instead of a single-task learning (STL) BERT model improved model performance while reducing model size, while BERT models performed slightly better than DistillBERT-based models. Sha et al. (2021) compared several traditional ML algorithms with fine-tuned bi-directional LSTM, CNN-LSTM, and single classification layer (SCL) models coupled with BERT on two large forum post datasets. The CNN-LSTM model slightly outperformed all other models. As the performance increase was marginal, the authors concluded that even using a simpler approach, such as SCL, coupled with BERT delivers a satisfactory classification performance.

Four papers (Yang et al., 2021; Baral et al., 2023; Srivastav et al., 2024; Zhang et al., 2022) utilized GenAI for **automated scoring or grading**. The data used in this group of papers varied from student essays and math problems to e-book learning materials. All papers used BERT models to build an automated scoring algorithm except for Baral et al. (2023), who implemented three different CLIP models by OpenAI, as some of the mathematics problems needed to be extracted from handwritten responses. The effectiveness of custom GenAI approaches in each was assessed against traditional algorithms, such as logistic regression, and/or other implementations of GenAI algorithms, such as SBERT-Canberra. The most used evaluation metrics were Cohen's κ and F1-score. BERT models were found to be the best-performing for scoring tasks compared to traditional machine learning models, such as Support Vector Machine or Decision Tree, and CLIP models; however, Yang et al. (2021) reported only moderate human–AI agreement and the inconsistent scoring performance for different groups of students.

Four papers (C. Li et al., 2022; Shin et al., 2022; Tsabari et al., 2023; Sarwat et al., 2022) used GenAI for **prediction**. Several different GenAI approaches were used in this group of studies. C. Li et al. (2022) used BERT to predict social support using discussion forum data. Tsabari et al. (2023) examined using a custom implementation of BERT, CodeBert, to predict "bug-fix-time" in programming assignments. Three papers aimed to predict student performance. A novel approach of using CGAN and SVM in Sarwat et al. (2022) showed higher sensitivity, specificity, and AUC than other methods, such as CNN, LSTM, or GGM. Shin et al. (2022) found promising results in GAN together with the Seldonian algorithm and path analysis to predict student performance, while trying to reduce bias.

Four papers (Corlatescu et al., 2024; C. Li et al., 2022; Phung et al., 2023; Shibani et al., 2023) integrated GenAI in their **tools**. All papers in this category used OpenAI GenAI models. Corlatescu et al. (2024) integrated ChatGPT in an updated version of a tool modelling reading comprehension, AMoC v4.0. Several validation studies showed that GenAI implementation improved tool performance. Shibani et al. (2023) developed CoAuthorViz, a tool that allows writers to collaborate with GPT-3, captures the interaction data, and provides analytics of the human–AI collaboration. PyFiXV by Phung et al. (2023) utilizes Codex models to generate feedback on programming assignments. In their study, different feedback generation techniques were evaluated based on their precision and coverage. Finally, C. Li et al. (2022) presented SafeMathBot built on GPT-2. The goal of this bot is to increase safe interactions in discussion forums. SafeMathBot was found to be more effective in increasing discourse safety than two other GenAI bots: BlenderBot and DialoGPT.

Table 3. Reviewed Articles in the *Methods, Analysis, and Tools* Category

Ref.	Role of GenAI	Overall LA task	GenAI technology	Data type/sources	Evaluation method	Results
J. Lee & Koh, 2023	Automated classification	Formative assessment of student teamwork	BERT	Chat messages	Comparison with RF+TD-IDF; Baseline: Human coding; Evaluation metrics: P, R, F1, Hamming distance, Cohen's κ	RF had higher precision, but BERT performed better in other metrics and achieved higher Cohen's κ on unseen data
Y. Li et al., 2022	Automated classification	Learning objectives classification	BERT	Learning objectives	Comparison with NB, LogR, SVM, RF, XG-Boost; Baseline: Human coding; Evaluation metrics: A, Cohen's κ, AUC, F1	BERT models performed best
Jia et al., 2021	Automated classification	Peer assessment evaluation	STL-BERT, MTL-BERT, STL-DistilBERT, MTL-DistilBERT	Peer review comments	Comparison with STL-GloVe; Baseline: Human coding; Evaluation metrics: F1, AUC	STL-BERT outperformed STL-GloVe and reduced the need for labelled data; MTL-BERT performed best among all BERT models but high memory resource usage and response time
Hu et al., 2024	Automated classification	Automated feedback generation	DistilBERT+ Integrated Gradients method	MOOC discussion data	Comparison with RF (previous research results); Baseline: Human coding; Evaluation metrics: F1	DistillBERT performed best
Z. Liu et al., 2023	Automated classification	Detecting cognitive presence in discussion forums	MOOC-BERT	MOOC discussion data	Comparison with DPCNN, FastText, TextRNN, AttBiLSTM, BERT (Chinese version); Evaluation metrics: A, P, R, F1	MOOC-BERT performed best
Samadi et al., 2024	Automated classification	Modelling of collaborative problem-solving skills	BERT, GPT-2 models	Utterance data	Comparison with RF, SVM, NB, RNN, CNN; Baseline: Human coding; Evaluation metrics: A, F1	BERT performed best
Sha et al., 2021	Automated classification	Forum post classification	BERT+CNN-LSTM, BERT+Bi-LSTM	Forum posts	Comparison with NB, SVM, RF, LR; Evaluation metrics: A, Cohen's κ, AUC, F1	Fine-tuned BERT+CNN-LSTM model performed best, but limited improvement; Integrating BERT at the classification layer increased the performance of CNN-LSTM and Bi-LSTM models
Srivastav et al., 2024	Automated scoring	Automated feedback generation	BERT	Student essays	Comparison with SVM, NB, LR, RF, DT, KNN, CNN, LSTM, GRU, PARs-BERT, RoBERT, and SemBERT; Evaluation metrics: P, R, F1	BERT-based model performed best
Zhang et al., 2022	Automated scoring	Assessment of short math answers	MathBERT	Student responses to open-ended questions	Comparison with SBERT-Canberra; Baseline: Human coding; Evaluation method: Rasch model; Evaluation metrics: AUC, RMSE, Cohen's κ	MathBERT performed best
Baral et al., 2023	Automated scoring	Prediction of scores for student responses	CLIP-Text, CLIP-Image, CLIP-OCR	Open-ended mathematics problems and scores (including handwritten responses)	Comparison with SBERT-Canberra, RF, Evaluation method: Rasch model; Evaluation metrics: AUC, RMSE, Cohen's κ	SBERT-Canberra performed best

Yang et al., 2021	**Automated scoring**	Modelling e-book text marking skills	BERT	Accounting e-book learning materials	Comparison with human scoring; Evaluation metrics: Cohen's κ, P, R, F1	Moderate human–AI agreement; BERT graded low-performing students more correctly than high-performing students
Sarwat et al., 2022	**Prediction**	Performance prediction	cGAN+SVM (multiple kernel types)	Student performance records, log data, questionnaire data	Comparison with RF, LR, ETC, GBM, SGD, CNN, LSTM; Evaluation metrics: Sensitivity, Specificity, AUC	CGAN+SVM performed best
Tsabari et al., 2023	**Prediction**	"Bug-fix-time" prediction	CodeBert	Code submission instances	Comparison with Halstead Metric Based Method and Critical code only; Baseline: Code2Vec Based Method; Evaluation metrics: AUROC, R, F1	CodeBert performed best
Shin et al., 2022	**Prediction**	Bias in predictive learning analytics	GAN + Seldonian algorithm+path analysis	Log data, surveys	Evaluation metrics: A, F1, AUROC	GAN+Seldonian algorithm+path analysis effective in reducing bias in prediction
C. Li et al., 2022	**Prediction**	Social support prediction	BERT	Discussion forum posts	Comparison with SVM, DT, RF; Evaluation metrics: MAE, MSE	BERT performed best
Corlatescu et al., 2024	**Tool development:** AMoC v4.0	Modelling reading comprehension	ChatGPT	The activation matrix from the original Landscape Model	Validation studies comparing with previous tool versions	AMoC v4.0 with ChatGPT integration outperforms previous versions
Shibani et al., 2023	**Tool development:** CoAuthorViz	Human–AI writing collaboration	GPT- 3	Creative stories, argumentative essays, keystroke-level logs	TAACO writing indicators	-
Phung et al., 2023	**Tool development:** PyFiXV	Generating feedback for programming syntax errors	Codex: code-davinci-edit-001 code-davinci-002	Programming assignments	Comparison of different feedback generation techniques; Evaluation metrics: F1, P, C, Qualitative analysis	High precision feedback through a run-time validation mechanism
C. Li et al., 2022	**Tool development:** SafeMathBot	Responsible conversational agents	SafeMathBot (GPT-2)	Discussion forum posts	Comparison with BlenderBot and DialoGPT	SafeMathBot can effectively enhance the safety

A: Accuracy; AUROC: Area under the ROC curve; C: Coverage; CGAN: Conditional GAN; CNN: Convolutional Neural Network; DT: Decision Tree; ETC: Extra Tree Classifier; GBM: Gradient Boosting Machine; GloVe: Global Vectors for Word Representation; GRU: Gated recurrent unit; KNN: K-Nearest Neighbour; LSTM: Long Short-Term Memory; LR: Logistic Regression; MAE: Mean Absolute Error;, MSE: Mean Square Error; NB: Naïve Bayes; P: Precision; R: Recall; RF: Random Forest; RMSE: Root Mean Square Error; RNN: Recurrent Neural Network; SGD: Stochastic Gradient Descent; SVM: Support Vector Machine.

4.4. Implementation and Action

Implementation and action refers to operationalizations of GenAI and LA in a classroom, whereby students and/or teachers get to interact with the GenAI (see Table 4). We see a change in technologies used at this stage of the LA/EDM cycle, where commercial tools such as ChatGPT (J. Liu et al., 2024; X. Lin et al., 2024; Nguyen et al., 2024) and Midjourney (Milesi et al., 2024) are operationalized. Only one article used GANs (Jin et al., 2023), and two studies used other tools based on GPT (J. Liu et al., 2024; Pinargote et al., 2024) or Stable Diffusion (U. Lee et al., 2023).

The largest subcategory is that of **human–GenAI collaboration**, which has five papers (J. Liu et al., 2024; X. Lin et al., 2024; Nguyen et al., 2024; Nguyen et al., 2024; Jin et al., 2023; U. Lee et al., 2023). In these studies, students used GenAI tools to complete learning tasks in diverse contexts, ranging from writing tasks (J. Liu et al., 2024; Nguyen et al., 2024), forum discussions (X. Lin et al., 2024), drawing (Jin et al., 2023), or generating images (U. Lee et al., 2023). The interactions between students and GenAI have been studied from a temporal lens (J. Liu et al., 2024; Nguyen et al., 2024) using established LA methods such as Lagged Sequence Analysis (LSA), Sequence Analysis (SA), HMMs, or Process Mining (PM). A prerequisite to being able to analyze the GenAI-supported learning process is to code the interactions between students and the AI tools (J. Liu et al., 2024; Nguyen et al., 2024). Other studies used more traditional statistical methods (X. Lin et al., 2024; Jin et al., 2023; U. Lee et al., 2023) or qualitative analysis (U. Lee et al., 2023). The findings showed that GenAI tools can increase

participation (X. Lin et al., 2024) and decision-making (Jin et al., 2023), and that different patterns of usage can be detected using LA methods (J. Liu et al., 2024; Nguyen et al., 2024; U. Lee et al., 2023).

The next subcategory is **LA dashboard development,** with two studies (Milesi et al., 2024; Pinargote et al., 2024), both of which used GenAI to present dashboard insights in a more accessible way. The work reported in Milesi et al. (2024) used GenAI image generation to create a comic-based story to present LA insights to students who participated in a nursing simulation. Students found data comics to be more engaging, enjoyable, and accessible than conventional visualizations, but potentially distracting from essential information in the context of higher nursing education. In turn, in Pinargote et al. (2024), the authors used GenAI to convert the dashboard insights into textual explanations. Students generally found metrics charts and feedback useful for understanding collaborative contributions, though opinions on their fairness and the necessity of detailed feedback varied.

Table 4. Reviewed Articles in the *Implementation and Action* Category

Ref.	Role of GenAI	Overall LA task	GenAI technology	Data type/sources	LA methods	Evaluation method	Results
J. Liu et al., 2024	**Human–GenAI collaboration**: Groupwork	Analysis of learning behaviour and performance while using GenAI to complete an assignment	ERNIE Bot (like ChatGPT 3.5)	Coded student–AI interactions, Pre–post test	LSA, ENA	Analysis of student–GenAI interactions through ENA and LSA comparing between achievement groups	High performers adhered to a structured application framework and exercised cognitive agency.
Milesi et al., 2024	**Dashboard development**	Presenting each student their own performance data solving a simulation through a dashboard	Midjourney (Niji Model v5)	Simulation data, AI-generated data storytelling comic books, interviews	Spatial analytics, SNA, ENA	Qualitative analysis of student interviews on their reflections of their own LA data presented through GenAI comic	Students found data comics to be more engaging, enjoyable, and accessible than conventional visualizations, particularly for visual learners, but criticized them for lacking professionalism and potentially distracting from key information in the context of higher nursing education.
X. Lin et al., 2024	**Human–GenAI collaboration:** Asynchronous online discussions	Analyzing student participation in GenAI-aided online discussions	ChatGPT	Log data of discussion boards	Descriptive statistics	Comparison of student participation before and after ChatGPT	Allowing ChatGPT on the discussion board significantly increased overall student participation
Nguyen et al., 2024	**Human–GenAI collaboration:** Academic writing tasks	Understanding student tactics and strategies when writing using GenAI	ChatGPT	Background pre-survey, Coded writing-related actions (screen capture), Task grade	SA, HMM, PM	Analyze the patterns doctoral students employ in their academic writing processes	Three main writing tactics arise: Content Pasting, Content Copying, and Component Shaping. The sequence of patterns gives rise to two writing strategies: Structured Adaptively and Unstructured Streamline
Pinargote et al., 2024	**Dashboard development**	Presenting students' GenAI-created insights about their learning data in a dashboard	Data Narratives Generator (based on GPT 3.5)	Students' coded collaboration summary data, Think-aloud, Questionnaire	Descriptive data visualization	Qualitative analysis of student exploration of the dashboard through think-aloud; Descriptive statistics of the questionnaire responses	Students generally found metrics charts and feedback useful for understanding collaborative contributions, though opinions on their fairness and the necessity of detailed feedback varied. Most students appreciated the clarity and fairness of the summaries provided, but some felt that assigning

							specific roles could be unfair or that in-person feedback was more meaningful.
Jin et al., 2023	**Human–GenAI collaboration:** Learning to draw	Analysis of students' drawing performance when aided by a GenAI system	GAN	Drawing completing time and difficulty level of subtasks	Descriptive statistics and data visualization	Analysis of student completion times and attempted difficulty levels	The GenAI-supported drawing tool helps improve student understanding of image structure and decision-making
U. Lee et al., 2023	**Human–GenAI collaboration:** Art-focused STEAM classes	Analysis of students' GenAI image creation performance	Dream Studio (based on Stable Diffusion)	Prompts, Generated pictures, Imaginative diaries	Qualitative analysis, Descriptive statistics, Correlation analysis	Analysis of student prompting indicators Correlation analysis between divergent-convergent thinking and student prompting	Students who change topics more often (divergent thinking) make more intensive use of the GenAI

LSA: Lag Sequential Analysis; ENA: Epistemic Network Analysis; SNA: Social Network Analysis; SA: Sequence Analysis; HMM: Hidden Markov Models; PM: Process Mining.

4.5. Evaluation

Evaluation indicates the use of LA to examine the integration of GenAI and provide insights about the process as a whole for future implementations, analyzing its impact on **academic achievement** as well as **student perceptions** (see Table 5). The seven studies under this category used traditional statistical methods to analyze the relationship between GenAI usage and performance using diverse study designs. For instance, in J. Liu et al. (2024) and X. Lin et al. (2024) the authors compared student performance before and after using ChatGPT. In Nguyen et al. (2024), student performance is compared between two groups that present different GenAI usage patterns. In Jin et al. (2023), Tack and Piech (2022), and Banihashem et al. (2024), performance is compared across a control group that does not rely on AI and one or more experimental groups that use some form of GenAI. In U. Lee et al. (2023), the quality of student writing and prompt quality are correlated with the quality of the pictures they generated using GenAI. Overall, the findings conclude that GenAI may be capable of improving student performance, but only when used properly. Students showed significant improvements in writing tasks, engagement in online discussions, and performance in creative tasks like pencil drawing. Most of the studies also analyze student perceptions of GenAI using questionnaires (J. Liu et al., 2024; X. Lin et al., 2024; U. Lee et al., 2023) and interviews (U. Lee et al., 2023). Students appreciated the role of GenAI in enhancing problem-solving, analytical skills, and written communication, despite facing some challenges with superficial content and over-reliance on AI-generated responses (X. Lin et al., 2024). Overall, the integration of GenAI improved critical thinking, and deeper knowledge construction (J. Liu et al., 2024; X. Lin et al., 2024) and reduced the gender gap in interest in art practices (U. Lee et al., 2023), indicating its valuable role in modern education.

Table 5. Reviewed Articles in the *Evaluation* Category

Ref.	Role of GenAI	Overall LA task	GenAI technology	Data type/sources	LA methods	Evaluation method	Results
J. Liu et al., 2024	**Using LA to analyze GenAI integration/effectiveness**	Evaluating student performance and perceptions while using GenAI to complete an assignment	ERNIE Bot (like ChatGPT 3.5)	Pre–post test, Frequency of interactions with GenAI, Questionnaire	t-test, Descriptive statistics	Comparison of pre–post test scores; Descriptive statistics from the questionnaire about student perceptions; Comparison of types of interactions between performance groups	Students improved their task using GenAI; Students have a positive perception of GenAI in completing instructional design tasks
X. Lin et al., 2024	**Using LA to analyze GenAI integration/effectiveness**	Evaluating student performance and perceptions while using GenAI to complete an assignment	ChatGPT	Pre–post test (Critical Thinking Scale), Questionnaires (CEQ, CIQ)	ANOVA, t-test	Comparison of student performance before and after ChatGPT; Descriptive statistics to examine the questionnaire results and qualitative analysis for examining students' open-ended responses	Students' critical thinking skills remained the same; Students had a positive experience with ChatGPT's integration into online discussions and acknowledged its potential to develop generic skills

Nguyen et al., 2024	**Using LA to analyze GenAI integration/effectiveness**	Evaluating the difference in performance among different GenAI usage patterns	ChatGPT	Task performance, Students' GenAI user type	t-test	Comparison of the effective of GenAI-supported writing patterns	High achievers use GenAI tools effectively to improve their writing process. Low achievers do not make full use of the GenAI tools or view it only as an additional resource
Jin et al., 2023	**Using LA to analyze GenAI integration/effectiveness**	Evaluating students' drawing performance when aided by a GenAI system	GAN	Drawing completing time and difficulty level of subtasks	Visual comparison	Quasi-experimental design comparing perceptions and drawing performance between students who used the GenAI drawing tool and non-users	Students in the experimental group completed each subtask faster and with greater ease compared to the control group; The overall performance of students in the experimental group was better than that of students in the control group; Both students and teachers indicated that the GAN-system was helpful and motivating
Tack & Piech, 2022	**Using LA to analyze GenAI integration/effectiveness**	Evaluating GenAI's educational dialogue capabilities	Blender, GPT-3	Educational dialogue between student and teacher, Blender and GPT-3	Descriptive statistics, Data visualization, ANOVA, Correlation analysis	Comparison of teacher vs. GenAI educational dialogues in terms of pedagogical ability	Blender outperforms both GPT-3 and human teachers in terms of understanding and expanding on a student's utterance; however, both agents lag behind human performance in effectively helping students
U. Lee et al., 2023	**Using LA to analyze GenAI integration/effectiveness**	Evaluation of students' GenAI image creation performance	Dream Studio (based on Stable Diffusion)	Semi-structured interviews, Survey	t-test, correlations analysis, Qualitative analysis, NLP	Thematic analysis of interview data; Comparison of student learning patterns between genders; Correlation analysis between picture creation outputs and on language proficiency	The disparity in how boys and girls perceive art classes diminishes or nearly disappears when GenAI is used as a teaching tool; Students' lexical diversity and writing score in their diaries correlated with picture scores
Banihashem et al., 2024	**Using LA to analyze GenAI integration/effectiveness**	Evaluating GenAI feedback quality	ChatGPT	Essay writing quality, Manually coded peer and AI feedback	Correlation analysis	Comparison of affective, cognitive, and constructive variables between peer and AI-generated feedback; Relationship between feedback quality and essay writing quality	Peers provided feedback of higher quality compared to ChatGPT; As the quality of the essay improves, ChatGPT tends to provide more affective feedback, while peers tend to provide less affective feedback

NLP: Natural Language Processing; CEQ: Course Experience Questionnaire; CIQ: Critical Incident Questionnaire.

5. Discussion

This systematic literature review of 41 studies aimed to provide an overview of empirical research utilizing GenAI and LA. To achieve this goal, we categorized GenAI implementations and attempted to map each implementation onto a different stage in the LA pipeline. This process led us to develop the GenAI in the LA/EDM cycle framework (see Figure 3), consisting of five stages: 1) data, 2) pre-processing and preparation, 3) methods, analysis, and tools, 4) implementation and action, and 5) evaluation.

Within the five studies about the use of GenAI for ***Data*** generation, we can see that GANs are the dominating GenAI technology. The most common use case was augmenting datasets to achieve better predictive performance, a goal that was not fully realized since other algorithms for synthetic data generation achieved better performance in most cases, and there was barely any improvement compared to using the unbalanced datasets otherwise. GANs did perform better in terms of privacy preservation and generating realistic datasets. It is worth noting that all but one of the articles in this category used GenAI for synthesizing tabular data (numeric and categorical data), whereas only one article used this technology for the generation of open-ended text (exam questions), where GANs outperformed other data generation mechanisms. Interestingly, none of the reviewed papers relied on LLMs for the task of synthetic data generation, which remains an area in need of further inquiry.

Furthermore, using GenAI to simulate — and generate data for — learning scenarios is still an untapped potential that could help conceptualize and understand different situations (e.g., conflict in collaboration). Other potential areas also include generating textual data, social networks, relational and interconnected datasets. Also, while multimodal data is expensive to generate and requires sophisticated hardware, GenAI could be extremely helpful in this regard and generate datasets that can be used for training, experimenting and simulated scenarios. Most importantly, the LA community — and the education community at large — would need to develop standardized and more comprehensive evaluation metrics to assess the quality, diversity, and educational utility of synthetic data, which would also help advance research on GenAI methods for synthetic data generation.

While LA relied extensively on LMS and other digitally generated log data, textual data remains an essential part of learning research that offers a nuanced understanding of how students learn, collaborate, or regulate — among others. Yet, analyzing textual data is an exhaustive process that requires lengthy human labour in terms of cleaning, coding, and verifying the coded data. Therefore, the issue of efficiently analyzing textual data was a dominant challenge in the ***data pre-processing and preparation*** category. In particular, GenAI was used frequently to automatically code discourse data using either BERT or GPT models. The research aims of the papers show that many researchers see a potential use of implementing GenAI in the preprocessing stage in the pipeline to ultimately enhance dashboards with text data or to automatically generate feedback. Although there are some promising results in comparison to traditional NLP approaches, existing evidence does not support using GenAI to automatically code text data without rigorous validation of the GenAI output. Hence, more research is needed to establish reliable GenAI approaches to code text data. Two papers using GenAI to summarize text highlight a need to develop new methods to evaluate the quality and accuracy of GenAI-generated summaries, especially at scale. As evaluating text coding is, in essence, a classification task with many available evaluation metrics and methods, text summarization presents a larger challenge, where humans may need to assess the summary quality. Finally, there is early evidence that GenAI embedding algorithms have the potential to improve the performance of established NLP methods. Notwithstanding the incentive work by researchers in harnessing AI for these tasks, we are far from there yet. However, given the fast pace of development of the landscape of LLM and AI at large, along with data pre-processing methods, it is expected that the future could be fruit-bearing for textual analytics and that LLMs could help with several exhaustive tasks.

The examination of ***methods, analysis, and tools*** papers showed mostly research focused on improving the performance of automated classifiers and automated scoring by utilizing BERT models. Although GenAI approaches were found to perform better than traditional machine learning methods, these results — like textual data analysis — do not meet the standards of human evaluation. Furthermore, most reported results were dependent on the contextual conditions, e.g., student characteristics. Using GenAI methods for prediction included several methods, e.g., BERT and GANs, and similarly to the case of scoring and classification applications, the GenAI solutions outperformed traditional machine learning models. These gains — while not spectacularly larger — are promising given the accelerating evolution of GenAI. Our included papers showed a variety of possible integrations of GenAI in tools for the purposes of tracking and modelling human–GenAI writing collaboration, modelling reading comprehension more effectively than previous versions of the tool without GenAI integration, generating high-precision feedback on programming assignments, and building a bot that effectively enhances the safety of interactions in a discussion forum. All these tools used GenAI by OpenAI, either GPT models or Codex, which was specifically trained for programming data. Given the limited generalizability of case studies, the diverse landscape of experiments, and the accelerating development of GenAI, it is too early to conclude which method was better or will continue to be so in other contexts. Nevertheless, it is clear that GenAI opens new possibilities for more precision in the automated assessment of student assignments and modelling of student behaviour. When more research is available, a consolidated view of what works in which scenarios will be clearer.

Whereas the previous categories — data generation or preparation — contain applications that may be considered auxiliary applications in the analytics pipelines, ***Implementation and action*** may be the most important phase for the application of GenAI. In fact, the enthusiasm behind **GenAI** was mostly driven by the idea that it can support teaching and learning and automate labour-intensive human tasks, especially the cognitively intensive ones, a goal that is shared by LA as a field, to optimize learning and teaching and the environment where it occurs. In our dataset, the ***Implementation and action*** category had seven papers that address either LA of GenAI or GenAI of LA. The former is represented by the study of **Human–GenAI collaboration** implementations in the classroom. The reviewed studies used LA methods to analyze the interactions between students and GenAI. This follows one of the main research strands in LA whereby several methods (SNA, PM, ENA, LSA, etc.) are used to analyze student interactions. As with the study of other types of collaboration, human–GenAI interaction requires coding of each utterance to enable a nuanced quantitative analysis, and so far, GenAI has not proven reliable in automating these coding tasks. As such, we are at the stage where the capabilities of GenAI are explored — or shall we say, exploited. Nevertheless, a considerable amount of human labour and research is still necessary to understand human–GenAI interactions as they evolve. Furthermore, given the recency of the concept of human–GenAI interaction, there is a lack of theoretical understanding of such relationships, and, consequently, no mature framework for coding or classifying these interactions. Such a framework may need to be context-dependent since student interactions do not stop at the human–GenAI

interface but rather continue in a different environment (e.g., assignment, writing, or discussion) that enables different types of behaviour (e.g., editing in the case of writing, executing in the case of programming code).

Our study followed the constant and prevalent theme of studies using GenAI in education, where most feedback revolved around writing-intensive tasks. Only a few instances of image generation exist, and no examples in other contexts, such as mathematics or programming, even though we have seen in the previous category that ad hoc tools have already been developed for particular contexts. As such, there are plenty of opportunities for further research in LA of GenAI in these new environments. For instance, no studies used NLP methods (or even GenAI itself) to automatically code student interactions with GenAI. Moreover, all articles have examined one-on-one interactions between students and GenAI, with no research on how GenAI could be integrated as a team member.

Regarding GenAI of LA — i.e., the use of GenAI to augment LA implementations — only two of the reviewed studies fall under that category and both cover the topic of **dashboard development**. One of the studies deals with generating images to enhance the delivery of LA insights, and the other uses GenAI to add explanations to a descriptive LA dashboard. Given the limited attention that this topic has gathered, there are plenty of opportunities for further research; for example, using GenAI as a way of explaining predictive analytics to students (in an Explainable AI fashion).

The ***Evaluation*** category includes seven articles in which the integration of GenAI was evaluated in terms of learning effectiveness and/or student perceptions. Articles in this category relied on classic study designs and methodologies commonly employed in education technology research (e.g., pre–post test designs or quasi-experimental designs), with few instances of the use of more modern LA methods. The findings paint an optimistic picture: GenAI positively impacts task performance and online participation when used appropriately. Moreover, students show positive perceptions toward the integration of GenAI and acknowledge its potential for skill improvement. Evidence on the effectiveness of GenAI on actual learning (rather than task performance) is still lacking, and further research is needed to determine best practices to maximize knowledge and skill acquisition while limiting over-reliance on GenAI.

Ethics is a crucial dimension of integrating GenAI and LA. Surprisingly, only a third of the papers included in our review discussed ethical considerations. Two papers mentioned the need to consider ethical challenges (U. Lee et al., 2023) or listed several ethical issues with GenAI, such as security risks and concerns through command injection, data poisoning, bias, and malicious content (Pinargote et al., 2024) when implementing GenAI, without offering any solutions. A call for institutional solutions through collaboration among different stakeholders to produce standardized ethical implementation guidelines was mentioned in three papers (Srivastav et al., 2024; Nguyen et al., 2024; J. Liu et al., 2024). In three papers, GenAI is proposed as a solution to existing ethical issues. Y. Li et al. (2022) discussed using synthetic data generation to ensure student privacy, while Ying and Ma (2024) and C.-C. Lin et al. (2024) mentioned the issues of students not consenting to data collection as a motivation to enhance datasets with synthetic data. Other papers suggested some ideas to ensure a more ethical implementation of GenAI. Baral et al. (2023) encouraged advanced data pre-processing methods and evaluation mechanisms of the GenAI performance to check for fairness and discouraged including sensitive variables in GenAI models, while Misiejuk, Kaliisa, and Scianna (2024) advised bias-mitigating practices being embedded throughout the implementation process of GenAI to address bias issues.

The rapid pace of GenAI development and the vastly surprising surge in LLM capabilities have no doubt taken the education community by surprise. A large diverse landscape of research has emerged to exploit the possibilities of GenAI and optimize learning and teaching. Nevertheless, knowledge building and research take time to settle and converge to established reliable applications. It is far too early to judge GenAI and its applications now or talk of any long-term effects. Relevant to this discussion is the diverse landscape of GenAI applications that were analyzed with methods borrowed from human–human and human–computer interactions. Future research may need to establish appropriate methods for the emerging and novel field of human AI work. Last, and most importantly, our systematic literature review has mapped the current landscape of GenAI in LA, but the future may be different. When the hype settles, we will know what GenAI is useful for and how, what may need human supervision, what should be avoided altogether, and what we need to do to improve.

This literature review provides several implications for practitioners, researchers, and developers:

***Implications for practitioners*:** The findings of this review highlight the potential of GenAI tools in educational contexts, particularly for educators. These tools can generate personalized learning materials tailored to individual student needs (Yusuf et al., 2024), which can save educators time in tasks such as grading and feedback provision. Moreover, the integration of GenAI tools like ChatGPT can enhance pedagogical practices through virtual simulations and multilingual support, especially in disciplines like language education and programming (Haindl & Weinberger, 2024). However, as noted in previous studies (e.g., Yu, 2024), successful adoption of these technologies requires educators to be well trained in their ethical and pedagogical applications. This necessitates the need for empirical evidence on the effectiveness of GenAI tools to justify the investment of resources into AI development and training of educators to support its use in teaching practice. Moreover, the integration of GenAI has implications for assessment. For instance, students' potential overreliance on GenAI begs the need for new forms of evaluation in which the role of AI in student learning is considered, focusing on students' higher-order thinking skills and self-regulation rather than on task performance only (Xia et al., 2024). Lastly, future research should explore the effects of

GenAI integration across educational contexts, particularly in understudied areas like K–12 education and culturally diverse settings (Yusuf et al., 2024). Expanding research to address these gaps will inform policy and practice and justify investment in GenAI technologies.

Implications for researchers: The review highlights the potential of GenAI to support multimodal LA by synthesizing different data types such as text, speech, and visual inputs (Yusuf et al., 2024) and supporting researchers to code large volumes of text (Misiejuk, Kaliisa, & Scianna, 2024). These applications could save researchers time and enable them to scale their analyses across larger datasets and complex multimodal environments. For example, studies have shown that GenAI can align well with predefined coding schemes, making it particularly useful for repetitive tasks in data preprocessing and analysis. However, as Misiejuk, Kaliisa, and Scianna (2024) cautioned, researchers should be mindful of GenAI's limitations in contextual interpretation, which may result in oversimplified or inaccurate coding outputs. Therefore, researchers are encouraged to employ hybrid methods that combine the efficiency of GenAI with the interpretative depth of human oversight. Additionally, while the generation of synthetic datasets can address data access challenges, researchers must validate these datasets to ensure accuracy and alignment with ethical requirements. On another vein, the examination of student interactions with GenAI in learning contexts remains largely unexplored with only five related studies. This task poses several challenges including data collection, coding of the student–AI interactions, as well as finding a suitable analytical lens that can account for the personalized nature of AI (e.g., idiographic analysis; Saqr et al., 2024). However, research advances in this line are needed to increase our understanding of how students engage with — and ideally benefit from — GenAI tools.

Implications for developers: Studies in this review have pointed out technical and ethical challenges associated with using GenAI in learning settings (Giannakos et al., 2024). Strategies to overcome such challenges include considering privacy-preserving techniques, such as federated learning and differential privacy, to address concerns about data security and misuse (Crompton & Burke, 2024). Moreover, participatory research and development approaches involving stakeholders (e.g., educators, students, and policymakers) can ensure that GenAI applications meet diverse educational needs while mitigating risks related to algorithmic bias and over-reliance on automated systems.

6. Limitations

This systematic review presents a comprehensive overview of empirical research integrating GenAI and LA. However, several limitations should be considered. As shown in the work by García-Peñalvo and Vázquez-Ingelmo (2023), the term *generative AI* can potentially encompass a variety of methods and approaches depending on a definition. In this paper, we filtered the papers in our dataset based on the methods widely recognized as GenAI, such as transformers or GANs; however, other conceptualizations may be feasible. Similarly, the boundaries of the fields of LA and EDM are not sharp, as both use methods and approaches also utilized in other research fields. At the same time, some researchers may not mention LA or EDM in their papers, although they may position themselves within either research field. As such, some work may not have been captured in our dataset if the authors did not explicitly mention LA or EDM in their papers or did not publish in a venue associated with either field. As with any literature review, the inclusion of papers is limited to the results retrieved from the databases based on our search string. Finally, we recognize the challenging process of classifying GenAI implementations in various stages of the LA pipeline, as well as in different subcategories. For example, BERT models are typically used for word embeddings, as described in various papers in our dataset. However, the extent of the importance of this task for GenAI implementation varies in each paper. In Menzel et al. (2023), BERT replaced Latent Semantic Analysis for word embeddings in a well-established analysis method in a novel approach, in Sha et al. (2021), models using BERT at the embedding layer were compared with BERT fine-tuning, while in Srivastav et al. (2024), the coupling of BERT embedding is an integral part of an automated feedback assessment model. This resulted in different classifications for similar GenAI implementations.

7. Conclusion

A primary goal of the fields of LA and EDM is to optimize learning and teaching by gaining insights into the learning process and learner behaviour. This systematic literature review showed the potential of GenAI to support these goals at each stage of the LA pipeline. GenAI was used for synthetic data generation for teaching and learning, to balance datasets or to ensure privacy, as well as for audio transcription, text summarization, and automated discourse coding. Moreover, we found several implementations of GenAI to support automated classification, automated scoring, and prediction. GenAI was successfully integrated into tools and LADs. Several studies explored the possibilities of human–GenAI collaboration in tasks ranging from writing to drawing. Finally, our review showed examples of studies that used LA methods to analyze GenAI integration and effectiveness. Although there is still a need to develop best practices of reliable GenAI integration into the LA pipeline and classroom teaching and learning practices, this review highlights that GenAI has already become another tool in the toolbox for LA methods.

A large, varied landscape of studies has tried to take advantage of the potential capabilities of GenAI in LA. An encouraging repertoire of applications has been reported across diverse domains. Yet, most applications are at the exploitation-exploration

stage, where researchers are trying to embed GenAI or replace existing tasks with GenAI capabilities. GenAI has been mostly useful in its generative capabilities and less frequently so for its cognitive capabilities that would enable researchers — or practitioners — to use it reliably for, e.g., teaching tasks, or to replace labour-intensive tasks such as coding text. Nonetheless, it is a fast-developing area of inquiry and GenAI is constantly gaining functionalities and expanding. It is, therefore, rather likely that these limitations will not stand the test of time, and GenAI may overcome them. Maybe the question is when — rather than would — this may happen and whether it would happen with current AI technologies (LLMs), or do we have to wait for another generation of AI?

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The paper is co-funded by the Academy of Finland (Suomen Akatemia) Research Council for the project "Towards Precision Education: Idiographic Learning Analytics (TOPEILA)," Decision Number 350560, which was received by Mohammed Saqr, and the project "Optimizing Clinical Reasoning in Time-Critical Scenarios: A Data-Driven Multimodal Approach (CRETIC)," Decision Number 360746, which was received by Sonsoles López-Pernas.

References

Bozkurt, A. (2024). Why generative AI literacy, why now and why it matters in the educational landscape? Kings, queens and GenAI dragons. *Open Praxis*, *16*(3), 283–290. https://doi.org/10.55982/openpraxis.16.3.739

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. In S. Dawson, C. Haythornthwaite, S. Buckingham Shum, D. Gašević, & R. Ferguson (Eds.), *LAK '12: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 134–138). ACM Press. https://doi.org/10.1145/2330601.2330636

Crompton, H., & Burke, D. (2024). The educational affordances and challenges of ChatGPT: State of the field. *TechTrends*, *68*(2), 380–392. https://doi.org/10.1007/s11528-024-00939-0

Dorodchi, M., Al-Hossami, E., Benedict, A., & Demeter, E. (2019). Using synthetic data generators to promote open science in higher education learning analytics. In R. Barga & C. Zaniolo (Eds.), *2019 IEEE International Conference on Big Data (Big Data 2019)* (pp. 4672–4675). IEEE. https://doi.org/10.1109/bigdata47090.2019.9006475

Ferguson, R. (2019). Ethical challenges for learning analytics. *Journal of Learning Analytics*, *6*(3), 25–30. https://doi.org/10.18608/jla.2019.63.5

García-Peñalvo, F. J., & Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(4), 7–16. https://doi.org/10.9781/ijimai.2023.07.006

Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., Järvelä, S., Mavrikis, M., & Rienties, B. (2024). The promise and challenges of generative AI in education. *Behaviour & Information Technology*, 1–27. https://doi.org/10.1080/0144929X.2024.2394886

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, *11*(12), Article 5467. https://doi.org/10.3390/app11125467

Guo, S., Zheng, Y., & Zhai, X. (2024). Artificial intelligence in education research during 2013–2023: A review based on bibliometric analysis. *Education and Information Technologies*, *29*(13), 16387–16409. https://doi.org/10.1007/s10639-024-12491-8

Gupta, P., Ding, B., Guan, C., & Ding, D. (2024). Generative AI: A systematic review using topic modelling techniques. *Data and Information Management*, *8*(2), Article 100066. https://doi.org/10.1016/j.dim.2024.100066

Haindl, P., & Weinberger, G. (2024). Does ChatGPT help novice programmers write better code? Results from static code analysis. *IEEE Access*, *12*, 114146–114156. https://doi.org/10.1109/ACCESS.2024.3445432

Hopfenbeck, T. N., Zhang, Z., Sun, S. Z., Robertson, P., & McGrane, J. A. (2023). Challenges and opportunities for classroom-based formative assessment and AI: A perspective article. *Frontiers in Education*, *8*, Article 1270700. https://doi.org/10.3389/feduc.2023.1270700

Johnson, N., & Phillips, M. (2018). Rayyan for systematic reviews. *Journal of Electronic Resources Librarianship*, *30*(1), 46–48. https://doi.org/10.1080/1941126X.2018.1444339

Kaliisa, R., Kluge, A., & Mørch, A. I. (2022). Overcoming challenges to the adoption of learning analytics at the practitioner level: A critical analysis of 18 learning analytics frameworks. *Scandinavian Journal of Educational Research*, *66*(3), 367–381. https://doi.org/10.1080/00313831.2020.1869082

Kumar, S., Musharaf, D., Musharaf, S., & Sagar, A. K. (2023). A comprehensive review of the latest advancements in large generative AI models. In R. N. Shaw, M. Paprzycki, & A. Ghosh (Eds.), *Advanced communication and intelligent systems: Second international conference, ICACIS 2023, Warsaw, Poland, June 16–17, 2023, revised selected papers, part I* (pp. 90–103). Springer Cham. https://doi.org/10.1007/978-3-031-45121-8_9

Lodge, J. M., de Barba, P., & Broadbent, J. (2023). Learning with generative artificial intelligence within a network of co-regulation. *Journal of University Teaching and Learning Practice*, *20*(7), Article 02. https://doi.org/10.53761/1.20.7.02

Luckin, R., Cukurova, M., Kent, C., & du Boulay, B. (2022). Empowering educators to be AI-ready. *Computers and Education: Artificial Intelligence*, *3*, Article 100076. https://doi.org/10.1016/j.caeai.2022.100076

McCalla, G. (2023). The history of artificial intelligence in education: The first quarter century. In B. du Boulay, A. Mitrovic, & K. Yacef (Eds.), *Handbook of artificial intelligence in education* (pp. 10–29). Edward Elgar Publishing. https://doi.org/10.4337/9781800375413.00010

Misiejuk, K., López-Pernas, S., Kaliisa, R., & Saqr, M. (2024). *Learning together: Modeling the process of student–AI interactions to generate learning resources.* ResearchGate. https://www.researchgate.net/publication/387864942_Learning_together_Modeling_the_process_of_student-AI_interactions_when_generating_learning_resources

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Ouyang, F., & Zhang, L. (2024). AI-driven learning analytics applications and tools in computer-supported collaborative learning: A systematic review. *Educational Research Review*, *44*, Article 100616. https://doi.org/10.1016/j.edurev.2024.100616

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W. Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, Article 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Perrotta, C., & Selwyn, N. (2020). Deep learning goes to school: Toward a relational understanding of AI in education. *Learning, Media and Technology*, *45*(3), 251–269. https://doi.org/10.1080/17439884.2020.1686017

Pressey, S. L. (1926). A simple apparatus which gives tests and scores-and teaches. *School and Society*, *23*, 373–376.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), Article e1355. https://doi.org/10.1002/widm.1355

Saqr, M. (2018). *Using learning analytics to understand and support collaborative learning* [Doctoral dissertation, Stockholm University]. DiVA. https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-159479

Saqr, M., Cheng, R., López-Pernas, S., & Beck, E. D. (2024). Idiographic artificial intelligence to explain students' self-regulation: Toward precision education. *Learning and Individual Differences*, *114*, Article 102499. https://doi.org/10.1016/j.lindif.2024.102499

Schöbel, S., Schmitt, A., Benner, D., Saqr, M., Janson, A., & Leimeister, J. M. (2024). Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers. *Information Systems Frontiers*, *26*(2), 729–754. https://doi.org/10.1007/s10796-023-10375-9

Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: A systematic review and applications. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-20016-1

Skinner, B. F. (1958). Teaching machines: From the experimental study of learning come devices which arrange optimal conditions for self-instruction. *Science*, *128*(3330), 969–977. https://doi.org/10.1126/science.128.3330.969

Suppes, P. (1966). The uses of computers in education. *Scientific American*, *215*(3), 206–223. https://doi.org/10.1038/scientificamerican0966-206

Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, *21*(1), Article 40. https://doi.org/10.1186/s41239-024-00468-z

Yan, L., Martinez-Maldonado, R., & Gašević, D. (2024). Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. In B. Flanagan, B. Wasson, & D. Gašević (Eds.), *LAK '24: Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 101–111). ACM Press. https://doi.org/10.1145/3636555.3636856

Yu, H. (2024). The application and challenges of ChatGPT in educational transformation: New demands for teachers' roles. *Heliyon*, *10*(2), Article e24289. https://doi.org/10.1016/j.heliyon.2024.e24289

Yusuf, A., Pervin, N., Román-González, M., & Noor, N. M. (2024). Generative AI in education and research: A systematic mapping review. *Review of Education*, *12*(2), Article e3489. https://doi.org/10.1002/rev3.3489

Zhan, C., Deho, O. B., Zhang, X., Joksimović, S., & de Laat, M. (2023). Synthetic data generator for student data serving learning analytics: A comparative study. *Learning Letters*, *1*, Article 5. https://doi.org/10.59453/KHZW9006

Appendix A: Papers Included in the Systematic Literature Review

[40] Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, *21*(1), Article 23. https://doi.org/10.1186/s41239-024-00455-4

[27] Baral, S., Botelho, A., Santhanam, A., Gurung, A., Cheng, L., & Heffernan, N. (2023). Auto-scoring student responses with images in mathematics. In M. Feng, T. Käser, & P. Talukdar (Eds.), *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 362–369). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115645

[30] Condor, A., Litster, M.& Pardos, Z. (2021) Automatic short answer grading with SBERT on out-of-sample questions. In I.-H. Hsiao, S. Sahebi, F. Bouchet, & J.-J. Vie (Eds.), *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 345–352). International Educational Data Mining Society.

[2] Corlatescu, D.-G., Watanabe, M., Ruseti, S., Dascalu, M., & McNamara, D. S. (2024). The automated model of comprehension version 4.0: Validation studies and integration of ChatGPT. *Computers in Human Behavior, 154*, Article 108154. https://doi.org/10.1016/j.chb.2024.108154

[1] Garg, R., Han, J., Cheng, Y., Fang, Z., & Swiecki, Z. (2024). Automated discourse analysis via generative artificial intelligence. In B. Flanagan, B. Wasson, & D. Gašević (Eds.), *LAK '24: Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 814–820). ACM Press. https://doi.org/10.1145/3636555.3636879

[13] Hu, Y., Giacaman, N., & Donald, C. (2024). Enhancing trust in generative AI: Investigating explainability of LLMs to analyse confusion in MOOC discussions. In B. Flanagan, A. Shimada, F. Okubo, H.-T. Tseng, A. C. M. Yang, O. H. T. Lu, & H. Ogata (Eds.), *Joint proceedings of LAK 2024 workshops co-located with 14th International Conference on Learning Analytics and Knowledge (LAK 2024)* (pp. 195–204). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3667/GenAILA-paper3.pdf

[31] Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., & Gehringer, E. (2021). All-in-one: Multi-task learning BERT models for evaluating peer assessments. In I.-H. Hsiao, S. Sahebi, F. Bouchet, & J.-J. Vie (Eds.), *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 525–532). International Educational Data Mining Society.

[22] Jin, Y., Li, P., Wang, W., Zhang, S., Lin, D., & Yin, C. (2023). GAN-based pencil drawing learning system for art education on large-scale image datasets with learning analytics. *Interactive Learning Environments*, *31*(5), 2544–2561. https://doi.org/10.1080/10494820.2019.1636827

[37] Lee, J., & Koh, E. (2023). Teamwork dimensions classification using BERT. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, Doctoral Consortium and Blue Sky. 24th international conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, proceedings* (pp. 254–259). Springer Cham. https://doi.org/10.1007/978-3-031-36336-8_39

[24] Lee, U., Han, A., Lee, J., Lee, E., Kim, J., Kim, H., & Lim, C. (2023). Prompt aloud! Incorporating image-generative AI into STEAM class with learning analytics using prompt data. *Education and Information Technologies*, *29*(8), 9575–9605. https://doi.org/10.1007/s10639-023-12150-4

[3] Li, C., Xing, W., & Leite, W. (2022). Building socially responsible conversational agents using big data to support online learning: A case with Algebra Nation. *British Journal of Educational Technology*, *53*(4), 776–803. https://doi.org/10.1111/bjet.13227

[38] Li, Y., Raković, M., Poh, B. X., Gašević, D., & Chen, G. (2022). Automatic classification of learning objectives based on Bloom's taxonomy. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 530–537). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853191

[16] Lin, C.-C., Cheng, E. S. J., Huang, A. Y. Q., & Yang, S. J. H. (2024). DNA of learning behaviors: A novel approach of learning performance prediction by NLP. *Computers and Education: Artificial Intelligence*, *6*, Article 100227. https://doi.org/10.1016/j.caeai.2024.100227

[8] Lin, X., Luterbach, K., Gregory, K. H., & Sconyers, S. E. (2024). A case study investigating the utilization of ChatGPT in online discussions. *Online Learning*, *28*(2). https://doi.org/10.24059/olj.v28i2.4407

[4] Liu, J., Li, S., & Dong, Q. (2024). Collaboration with generative artificial intelligence: An exploratory study based on learning analytics. *Journal of Educational Computing Research*, *62*(5), 1234–1266. https://doi.org/10.1177/07356331241242441

[39] Liu, Q., Khalil, M., Jovanović, J., & Shakya, R. (2024). Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics. In B. Flanagan, B. Wasson, & D. Gašević (Eds.), *LAK '24: Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 620–631). ACM Press. https://doi.org/10.1145/3636555.3636921

[12] Liu, Z., Kong, X., Chen, H., Liu, S., & Yang, Z. (2023). MOOC-BERT: Automatically identifying learner cognitive presence from MOOC discussion data. *IEEE Transactions on Learning Technologies*, *16*(4), 528–542. https://doi.org/10.1109/TLT.2023.3240715

[19] Menzel, L., Gombert, S., Weidlich, J., Fink, A., Frey, A., & Drachsler, H. (2023). Why you should give your students automatic process feedback on their collaboration: Evidence from a randomized experiment. In O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, & T. Papathoma (Eds.), *Responsive and sustainable educational futures: 18th European Conference on Technology Enhanced Learning, EC-TEL 2023, Aveiro, Portugal, September 4–8, 2023, proceedings* (pp. 198–212). Springer Cham. https://doi.org/10.1007/978-3-031-42682-7_14

[6] Milesi, M. E., Alfredo, R., Echeverria, V., Yan, L., Zhao, L., Tsai, Y.-S., & Martinez-Maldonado, R. (2024). "It's really enjoyable to see me solve the problem like a hero": GenAI-enhanced data comics as a learning analytics tool. In F. F. Mueller, P. Kyburz, J. R. Williamson, & C. Sas (Eds.), *CHI EA '24: Extended abstracts of the CHI Conference on Human Factors in Computing Systems* (Article 4). ACM Press. https://doi.org/10.1145/3613905.3651111

[36] Misiejuk, K., Kaliisa, R., & Scianna, J. (2024). Augmenting assessment with AI coding of online student discourse: A question of reliability. *Computers and Education: Artificial Intelligence*, *6*, Article 100216. https://doi.org/10.1016/j.caeai.2024.100216

[10] Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human–AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, *49*(5), 847–864. https://doi.org/10.1080/03075079.2024.2323593

[17] Pan, Z., Li, C., & Liu, M. (2020). Learning analytics dashboard for problem-based learning. In D. Joyner, R. Kizilcec, & S. Singer (Eds.), *L@S '20: Proceedings of the Seventh ACM Conference on Learning @ Scale* (pp. 393–396). ACM Press. https://doi.org/10.1145/3386527.3406751

[7] Phung, T., Cambronero, J., Gulwani, S., Kohn, T., Majumdar, R., Singla, A., & Soares, G. (2023). Generating high-precision feedback for programming syntax errors using large language models. In M. Feng, T. Käser, & P. Talukdar (Eds.), *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 370–377). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115653

[20] Pinargote, A., Calderón, E., Cevallos, K., Carrillo, G., Chiluiza, K., & Echeverria, V. (2024). Automating data narratives in learning analytics dashboards using GenAI. In B. Flanagan, A. Shimada, F. Okubo, H.-T. Tseng, A. C. M. Yang, O. H. T. Lu, & H. Ogata (Eds.), *Joint proceedings of LAK 2024 workshops co-located with 14th International Conference on Learning Analytics and Knowledge (LAK 2024)* (pp. 150–161). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3667/DS-LAK24_paper_5.pdf

[35] Pugh, S., Subburaj, S. K., Rao, A. R., Stewart, A. E., Andrews-Todd, J., & D'Mello, S. K. (2021). Say what? Automatic modeling of collaborative problem solving skills from student speech in the wild. In I.-H. Hsiao, S. Sahebi, F. Bouchet, & J.-J. Vie (Eds.), *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 55–67). International Educational Data Mining Society.

[28] Samadi, M. A., Jaquay, S., Lin, Y., Tajik, E., Park, S., & Nixon, N. (2024). Minds and machines unite: Deciphering social and cognitive dynamics in collaborative problem solving with AI. In B. Flanagan, B. Wasson, & D. Gašević (Eds.), *LAK '24: Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 885–891). https://doi.org/10.1145/3636555.3636922

[29] Sarwat, S., Ullah, N., Sadiq, S., Saleem, R., Umer, M., Eshmawi, A. A., Mohamed, A. & Ashraf, I. (2022). Predicting students' academic performance with conditional generative adversarial network and deep SVM. *Sensors*, *22*(13), Article 4834. https://doi.org/10.3390/s22134834

[32] Sha, L., Raković, M., Li, Y., Whitelock-Wainwright, A., Carroll, D., Gašević, D., & Chen, G. (2021). Which hammer should I use? A systematic evaluation of approaches for classifying educational forum posts. In I.-H. Hsiao, S. Sahebi, F. Bouchet, & J.-J. Vie (Eds.), *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 228–239). International Educational Data Mining Society.

[9] Shibani, A., Rajalakshmi, R., Mattins, F., Selvaraj, S., & Knight, S. (2023). Visual representation of co-authorship with GPT-3: Studying human–machine interaction for effective writing. In M. Feng, T. Käser, & P. Talukdar (Eds.), *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 183–193). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115695

[11] Shin, J., Bulut, O., & Pinto, W. N., Jr. (2022). E-learning preparedness: A key consideration to promote fair learning analytics development in higher education. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 673–678). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853111

[18] Snyder, C., Hutchins, N. M., Cohn, C., Fonteles, J. H., & Biswas, G. (2024). Analyzing students collaborative problem-solving behaviors in synergistic STEM+C learning. In B. Flanagan, B. Wasson, & D. Gašević (Eds.), *LAK '24: Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 540–550). ACM Press. https://doi.org/10.1145/3636555.3636912

[33] Srivastav, G., Kant, S., & Srivastava, D. (2024). Design of an AI-driven feedback and decision analysis in online learning with Google BERT. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(10s), 629–643. https://ijisae.org/index.php/IJISAE/article/view/4465

[23] Tack, A., & Piech, C. (2022). The AI teacher test: Measuring the pedagogical ability of Blender and GPT-3 in educational dialogues. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 522–529). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853187

[25] Tsabari, S., Segal, A., & Gal, K. (2023). Predicting bug fix time in students' programming with deep language models. In M. Feng, T. Käser, & P. Talukdar (Eds.), *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 396–405). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115733

[5] Vives, L., Cabezas, I., Vives, J. C., Reyes, N. G., Aquino, J., Cóndor, J. B., & Altamirano, S. F. S. (2024). Prediction of students' academic performance in the programming fundamentals course using long short-term memory neural networks. *IEEE Access*, *12*, 5882–5898. https://doi.org/10.1109/ACCESS.2024.3350169

[14] Volarić, T., Ljubić, H., Dominković, M., Martinović, G., & Rozić, R. (2023). Data augmentation with GAN to improve the prediction of at-risk students in a virtual learning environment. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial intelligence in education: Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, Doctoral Consortium and Blue Sky. 24th international conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, proceedings* (pp. 260–265). Springer Cham. https://doi.org/10.1007/978-3-031-36336-8_40

[15] Wu, Z., Deng, K., Qiu, J., & Tang, Y. (2022). ExamGAN and Twin-ExamGAN for exam script generation. *IEEE Transactions on Knowledge and Data Engineering*, 35(11), 11354–11367. https://doi.org/10.1109/TKDE.2022.3233046

[26] Wu, Z., Larson, E., Sano, M., Baker, D., Gage, N., & Kamata, A. (2023). Towards scalable vocabulary acquisition assessment with BERT. In D. Spikol, O. Viberg, A. Martínez-Monés, & P. Guo (Eds.), *L@S '23: Proceedings of the tenth ACM Conference on Learning @ Scale* (pp. 272–276). ACM Press. https://doi.org/10.1145/3573051.3596170

[21] Yang, A. C. M., Chen, I. Y. L., Flanagan, B., & Ogata, H. (2021). From human grading to machine grading. *Educational Technology & Society*, *24*(1), 164–175. https://www.jstor.org/stable/26977865

[41] Ying, D., & Ma, J. (2024). Student performance prediction with regression approach and data generation. *Applied Sciences*, *14*(3), Article 1148. https://doi.org/10.3390/app14031148

[34] Zhang, M., Baral, S., Heffernan, N. & Lan, A. (2022). Automatic short math answer grading via in-context meta-learning. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 112–132). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.6853032